What company do words keep?
Revisiting the distributional semantics of J.R. Firth & Zellig Harris

Mikael Brunila
McGill University
Department of Geography
mikael.brunila@gmail.com

Jack LaViolette
Columbia University
Department of Sociology
jack.laviolette@columbia.edu

Abstract

The power of word embeddings is attributed to the linguistic theory that similar words will appear in similar contexts. This idea is specifically invoked by noting that "you shall know a word by the company it keeps," a quote from British linguist J.R. Firth who, along with his American colleague Zellig Harris, is often credited with the invention of "distributional semantics." While both Firth and Harris are cited in all major NLP textbooks and many foundational papers, the content and differences between their theories is seldom discussed. Engaging in a close reading of their work, we discover two distinct and in many ways divergent theories of meaning. One focuses exclusively on the internal workings of linguistic forms, while the other invites us to consider words in new company—not just with other linguistic elements, but also in a broader cultural and situational context. Contrasting these theories from the perspective of current debates in NLP, we discover in Firth a figure who could guide the field towards a more culturally grounded notion of semantics. We consider how an expanded notion of "context" might be modeled in practice through two different strategies: comparative stratification and syntagmatic extension.

1 Introduction

> We are in the world and the world is in us.
>
> Alfred North Whitehead
> (1938; cited in Firth 1957c, 29)

If you have read any papers in computational linguistics in the past thirty years, you have likely come upon the following quote from British linguist J.R.Firth (1957c, 11): "You shall know a word by the company it keeps". Cited in most major textbooks (Manning and Schütze, 1999; Jurafsky and Martin, 2009; Eisenstein, 2019; Russell and Norvig, 2020), several foundational papers, and hundreds of other NLP articles, this phrase has come to index a theoretical orientation in a field that is increasingly focused on computation, often at the expense of linguistic theory (on these trends in NLP see Halevy et al., 2009; Manning, 2015; Norvig, 2012; Henderson, 2020; Church and Liberman, 2021). Together with American linguist Zellig Harris, Firth is regularly called upon to justify a distributional theory of semantics, whereby the meaning of lexical units is conceived in terms of relative co-occurrence and shared contexts of use (Sahlgren, 2008).

While Harris and Firth are often invoked, their ideas are seldom closely engaged. Hailing from disparate traditions, Harris and Firth had radically different ideas on the scope and context of linguistic analysis, and presented incongruent versions of the distributional method. Drawing on the information theory pioneered by Claude Shannon (1948), Harris was determined to work out a structuralist theory of language in terms of mathematical information (Léon, 2011; Nevin, 1993). Firth, on the other hand, came to linguistics via anthropology and borrowed heavily from pragmatic philosophies of language. For him, linguistic analysis always started with the "context of situation" and necessarily accounted for non-verbal actors and objects (Firth, 1957c, 9).

Considering the definition and extent of linguistic context is important for many reasons, as a spate of recent publications suggests (Glenberg and Robertson, 2000; Hovy, 2018; Bender and Koller, 2020; Bisk et al., 2020; Tamari et al., 2020; Trott et al., 2020). Firstly, it touches upon the limits of current paradigms in NLP, where corpus linguistics is perfected through increasingly complex models trained on increasingly massive corpora (Bender et al., 2021). This approach may advance the identification of linguistic *form*, but might ultimately have little to say about the relation of *meaning* to the social world (Bender and Koller, 2020; Bisk et al., 2020). Secondly, even with more modest ambitions, several NLP applications—e.g., with

spatial (McKenzie and Adams, 2021) or historical (Kutuzov et al., 2018) data—require that linguistic patterns be related to other types of structure. Thirdly, from sociological and sociolinguistic perspectives, meaning intrinsically varies as language is used in different settings and indexed to different social categories (Labov, 1972; Bourdieu, 1984; Silverstein, 2003; Eckert, 2008; Hovy, 2018). Finally, without a broader sense of context, NLP and language modeling in particular remains trapped in a paradigm where language is always treated as universal, making invisible both different speech communities (e.g. Nguyen et al., 2021) and the biases of language (e.g. Blodgett et al., 2020; Lu et al., 2020; Sap et al., 2020).

After a brief history of distributional semantics, we outline Harris and Firth's research on distribution and, more broadly, on the scope of linguistic analysis. We look at some of the ways in which NLP has tried to account for broader context within the distributional paradigm. We suggest that existing strategies can be understood in terms of either "comparative stratification" or "syntagmatic extension." We conclude with thoughts on why rereading Harris and, in particular, Firth might aid the field of NLP with its current aporias. If words shall be known by the company they keep, then the question follows: what *kind* of company do they keep? Are they found only alongside linguistic elements, or do they mingle with other types of entities? Or, as Firth himself wrote: "Many different answers could be given to the question 'Distribution of what, where and how?"' (Firth, 1957a, v).

2 Background: Distributional Semantics and NLP

Distributional semantics has been an fundamental part of computational linguistics since the beginnings of the field, but in a discontinuous manner encompassing at least two distinct eras. Firstly, during the 1950s and 60s, Harris was integral to the mathematization of linguistics in the US after the Second World War (Rubenstein and Goodenough, 1965; Léon, 2021). Firth was skeptical of efforts to mechanize linguistics,[1] but he nonetheless consulted for some of the early work on machine translation at Cambridge Language Research Group (Léon, 2007, 410), which included Firth's pupil, M.A.K. Halliday (Léon, 2021, 144) and shortly later the NLP pioneer Karen Spärck Jones (Léon, 2021, 89). Naturally, others also contributed to this first wave of distributional thinking, including Shannon (1945; 1948) with what might be considered one of the first language models, Warren Weaver (1952) with an early proposal for distributional semantics, and Martin Joos (1950) with a statistical formulation of language as a symbolic system of conditional probabilities.

Secondly, when computational linguistics returned to its "empiricist" roots in probabilistic methods and information theory in the mid-80s and early 90s (Norvig 2012; Léon 2021, 141), Firth and Harris accompanied Shannon among the authors who were invoked, in an ACL "Special Issue on Computational Linguistics Using Large Corpora," as foundational figures of a tradition that had been overshadowed for decades by the "rationalism" of characters such as Noam Chomsky and Marvin Minsky (Church and Mercer, 1993, 15). During this "corpus turn," the rapid automation of linguistics was driven by a resumed connection with postwar computational linguistics and information theory (Léon, 2021, 3). However, the 1990s wave of vector semantics papers that used methods like singular-value decomposition (SVD) to produce early "dense vector" models of meaning like LSA (Deerwester et al., 1989, 1990; Landauer and Dumais, 1997) and its derivatives (Hofmann, 1999; Blei et al., 2002), HAL (Burgess, 1998), or the models of Schütze (Schütze, 1992, 1993; Schütze and Pedersen, 1993) generally did not cite Firth or Harris, although a few papers from that period did (Church and Hanks, 1989; Hindle, 1990). In short, while Firth and Harris were not regularly used as stand-ins for linguistic theory during the 1990s and early 2000s, a general revival of empiricism and distributional approaches to meaning signaled a potential resurgence of interest in their thinking.

During the 2000s, the application of neural networks to language modeling tasks (e.g. Bengio et al., 2003) and the development of self-supervision techniques (e.g. Raina et al., 2007) set the stage for the word embedding breakthroughs of the early 2010s (e.g. Mikolov et al., 2013). By the end of the decade, the introduction of attention (Graves et al., 2013; Bahdanau et al., 2015) and then of the Transformer model (Vaswani et al., 2017) made way for the next breakthrough, the

[1] He seemed to consider the idea Orwellian (Firth, 1957b) and repeatedly attacked Norbert Wiener (e.g. Firth, 1968a,c), a pioneer whose work would later be considered foundational for connectionist AI (Goodfellow et al., 2016; Russell and Norvig, 2020).

large-language modeling revolution (e.g. Peters et al., 2018; Devlin et al., 2019; Chowdhery et al., 2022).

Following the introduction of the word2vec model and its powerful but "static" embeddings, Harris in particular was frequently cited (Le and Mikolov, 2014; Levy and Goldberg, 2014; Levy et al., 2015; Bojanowski et al., 2017), often (but not always) along with Firth (Bruni et al., 2014; Hamilton et al., 2016; Goldberg, 2017; Eisenstein, 2019; Jurafsky and Martin, 2021). However, despite an explosion of citations (Bisk et al., 2020, 8719), this interest has not been very engaged. In fact, the canonization of Firth and Harris during this time is paradoxical. On the one hand, it seems that they are invoked to lend theoretical authority to a field that struggles to lift its gaze from the latest state-of-the-art numbers (Manning, 2015; Bender and Koller, 2020). Yet, the unspoken conclusion from the ascent of neural models and the language modeling revolution was that "learning from data made linguistic theories irrelevant" (Henderson, 2020, 6295). In other words, just as NLP seemed to lose interest with linguistic theory, it elevated two pioneering theoreticians to canonical status, but seemingly without engaging closely with their work. In fact, it often seems as if Firth and Harris are referenced in such a cavalier manner that it deflects attention from the field's general lack of engagement with linguistic theory. Meanwhile, Firth and Harris became figures who justify a relatively narrow conception of meaning, one that is predominantly intra-linguistic, without much to say about its usage in social life.

This peculiar story has not been properly told. Though Léon discusses the contrast between Harris and Firth in the context of corpus linguistics (2008) and their influence on the history of computational linguistics (2021), her work does not address the differences in their distributional theories and conceptions of "context," nor the renewed and paradoxical significance of the two authors for language modeling. In our contribution, we emphasize the gap between the ideas of Firth and Harris as well as the insights a re-reading of their work offers for expanding the scope of computational semantics.

3 Harris's distributional structuralism

Few linguists contributed more to linguistic theory than Zellig Harris (1909–1992), and not just by serving as Noam Chomsky's doctoral advisor. In fact, the two came to share little in common (Goldsmith, 2005; Nevin, 2010). Whereas Chomsky's generative grammar repositioned linguistics as a cognitive science seeking to understand, in so few words, the idealized mental representations and structures enabling language acquisition and production (e.g. Chomsky, 1972), Harris' radically distributional approach to language effectively elevated the natural language corpus as the sole starting point from which linguistic theory could arise (Harris, 1951, 1; Harris et al., 1988, 2–3; Johnson, 2002, 143–144).

This theory consisted of a linguistic structure segmentable into a finite set of formal objects characterized by constrained patterns of correspondence with one another (Harris, 1951, 1954, 1991). Such patterns of correspondence can be observed only in language-in-use, that is, in natural language corpora. In his foundational paper "Distributional Structure," for example, Harris (1954, 156–157) provides a purely distributional account of how one might induce the semantic meanings of *oculist*, *eye-doctor*, and *lawyer* from the partial (in the case of *oculist* and *lawyer*) or nearly complete (in the case of *oculist* and *eye-doctor*) overlap in their observed "environments" of use. This approach applies to other levels of linguistic analysis, such as morphophonemics (e.g., Harris, 1954, 155). Rather than producing a series of descriptive rules for the distribution of each phoneme, morpheme, or word, greater parsimony was sought by grouping these elements into structurally equivalent classes—categorized by their relationships as "operators" and "arguments" in Harris' later work (e.g. 1968; 1988; 1991)—sharing the same distributional rules, compounding elements in a hierarchical manner.

In this section we draw attention to three aspects of Harris's distributional linguistics: the relationship it posits between meaning and form; assumptions about heterogeneity among speakers of the same dialect; and the concept of sublanguages.

3.1 Meaning and form

A result of Harris' vision of a linguistics—concerned above all with the structural and probabilistic constraints governing the combination of formal elements—is that the discipline would be fully autonomous, not only from biology and psychology but even from semantics, phonetics, and logic, "complete without intrusion of other features

such as history or meaning" (Harris 1954, 146; Goldsmith 2005, 725–726).[2] Harris's reasoning depended on the particular status he gave to linguistics among all the sciences. Taking language as its object of inquiry, linguistics lacks—unlike other sciences—a *metalanguage* external to language, i.e., to its object of inquiry (Harris, 1991, 4–5; Nevin, 1993, 356). Even if some other symbolic system is used, "those symbols will have to be defined ultimately in a natural language" (Harris, 1991, 274), as the surging demand for interpretability and explainability in NLP has made evident (see e.g. Danilevsky et al., 2020). Language is consequently not a "code" of "forms" that correlate with some meanings outside of it. It has no "one-to-one" conformity "with some independently discoverable structure of meaning" (Harris, 1954, 152). Instead, it is a system related to, but also independent from, thought (Harris, 1991, 383–384; Nevin, 1993, 361–363). While all human activity is meaningful, the particular meanings of language are *constituted* by its form, not *correlated* with it (Nevin, 1993, 394). Meaning thus understood is about departures from equiprobability in the distribution of these constraints (Harris, 1991, 23). These departures "define a range of meaning for each morpheme, which includes its meaning in each occurrence". Nonetheless, shared environments do not necessarily imply shared meanings: "bumped into a pole can be said after a minor accident or after a chance meeting with an East European" (Harris, 1951, 191).

The ultimate goal of Harris's linguistic inquiry is to evaluate the efficiency of different grammars and their ability to model the statistical constraints imposed upon the distribution of different linguistic elements (Goldsmith, 2005, 723–725). Harris held that language was a "detached pattern" (Harris, 1941, 295)—information that was public and socially transmissible and hence constitutive of new types of socially shared and conventional meaning (Harris, 1991, 342–345, 377–382; Nevin, 1993, 360, 365)—and linguistics could at best discover different incomplete grammars (Harris, 1991, 31–36). Though linguistics might provide insights about meaning and discourse, or about cultural practices, such findings would not bear directly on linguistics *per se* (Goldsmith, 2005, 725–726). Indeed, while Harris acknowledged that our sense of word meaning is aided by "extra-linguistic situational information," words "beyond the immediate situation" are "on their own" (Harris, 1991, 368). However, Harris's method is not completely detached. Searching for a method to segment speech, he notes that the similarity of elements "reduces ultimately to the similarity of sound segments under repetition," implemented through "the pair test" in which native speakers are asked to discriminate between sound segments (Harris, 1954, 158–159), producing an observational primitive that is "more easily controlled than data on meaning" (e.g. Harris, 1951, 20).

3.2 Variation, or lack thereof

Harris's view of language and linguistics, isolated from the vagaries of social interaction and variation, is obviously difficult to reconcile with a sociolinguistic perspective. As Harris writes in his *Structural Linguistics* (1951, 9), his approach is meant to describe a homogeneous dialect, which "[i]n most cases...presents no problem, since the whole speech of the person or community shows dialectal consistency." Referencing this passage, sociolinguist and dialectology pioneer William Labov (1966/2006, 5) argues that "the inconsistency found in most New York City idiolects is so great that the first alternative of Harris is impossible, and the second implausible."[3] In other words, even at the level of the speaker, Harris's idealized, unvarying idiolect does not hold up to empirical scrutiny. Rigorous consideration of factors that Harris would deem extra-linguistic (class, race, interactional roles, etc.) are indeed essential to produce a systematic description of linguistic structure (Labov, 1972). From the sociolinguistic perspective, Harris's vision of a linguistic science fully isolated from the "intrusion" of non-verbal social life would never obtain the systemicity to which it aspired.

This sociolinguistic critique highlights fundamental limitations of Harris's perspective. Language is viewed primarily through the distributional restrictions imposed by convention, rather than by "stylistic practice" and the ways in which speakers "make social-semiotic moves, reinterpreting variables and combining and recombining them in a continual process" (Eckert, 2012, 94). We can

[2] Indeed, as Jacqueline Léon notes (personal communication, April 18, 2022), even calling Harris' approach "semantics" is bit of an oxymoron. We elide a full discussion on the term since "distributional semantics" has become a commonplace phrase in NLP.

[3] New York City is not unique in this regard; it was merely the location of Labov's early pathbreaking work.

study changes in discourse, as Harris himself did in an impressive volume on structures in immunological theory over time (Harris et al., 1988), but not how people make those changes, or indeed the way in which it is "the variation itself that is systemic" (Deleuze and Guattari, 1987, 93).

3.3 Sublanguages

As noted above, Harris' revival in the 1990s was driven by the new interest in "corpus linguistics" of large corpora, a research paradigm that was partly derived from Harris' notion of "sublanguages" (Léon, 2021). Harris introduced sublanguages in his book *Mathematical Structures of Language*, defined as "[c]ertain proper subsets of a language [which] may be closed under some or all of the operations defined in the language, and [which] thus constitute a sublanguage of it" (Harris, 1968, 152). A sublanguage is a set of sentences which are a subset of the sentences of the "whole" language. However, the grammatical constraints of the sublanguage are not necessarily those of the whole language; rather, their grammars intersect (Harris 1968, Ch 11; Kittredge and Lehrberger 1982, 1). In application the term has come to refer primarily to the grammar and vocabulary unique to or characteristic of a particular professional or scientific field (e.g. Harris, 1988), an influential concept for early information retrieval research (Sager, 1975, 1981).[4] Harris believed that sublanguages could be neatly identified using the distributional methods of his general linguistic program.

4 Firth's contextual semantics

Something similar to the sociolinguistic critique of Harris could be articulated from a different perspective, namely, through the work of J.R. Firth. Firth (1890–1960)—professor of General Linguistics at the University College of London and the first holder of a chair in that subject in Britain—independently formulated a distributional theory of lexical semantics. However, unlike Harris, Firth refused to treat meaning separately from pragmatics, and words apart from their broader "context of situation." (Robins, 1997, 205–208)

Firth never published a fully articulated exposition of his general theory of language (Robins 1997, 216; Thomas 2011, 180) and today, all of his work is not only out of print but also mostly unavailable online. Not understood by "the contemporary scientism" of American descriptivist linguistics and its pioneers like Harris, Firth was mostly ignored on the other side of the Atlantic (Palmer 1968, 2; Pandit 1970, 280). Unlike many of his American contemporaries, Firth did not draw mainly from cognitive psychology and logic—the latter of which Firth thought had "taken the heart out of language" (Firth, 1957a, 186)—but from the work of Polish anthropologist Bronisław Malinowski (Robins, 1997, 211). Here, we focus on the evolution of his thoughts on meaning and collocation, as well as his notions of *context of situation* and *restricted language*.

4.1 Meaning by collocation

To Firth, the purpose of linguistics is to "study meaning in its own terms" (Firth 1968b, 145; Senis 2015, 289). The famous phrase about the company that words keep concerned a particular "mode of meaning": "meaning by collocation" (Firth, 1957b, 194). Anticipating vocabulary now ubiquitous in NLP, Firth thought that this level of meaning could be found by examining the "habitual collocations" of words and the "word-material" in which they are "most characteristically embedded" (Firth, 1957c, 11–12). Meaning by collocation was an abstraction of *syntagmatic relations* (Oyelaran 1967, 444) that went beyond "mere juxtaposition," stating instead "an order of *mutual expectancy*" and "mutual prehension" (Firth, 1957c, 12). While mutual expectancy could be understood similarly to Joos's (1950) conditional probabilities of occurrence or the concept of Pointwise Mutual Information (Fano, 1961), the notion of "prehension" originates in the work of philosopher and mathematician Alfred North Whitehead (1938, 1957; see also Butt 2013) and concerns the manner in which one entity grasps another and makes it part of its own experience (Christian 1959, 12; Bryant 2011, 136).

Drawing on Whitehead's "modes of thought" (1938; see also Butt 2001, 1812, Butt 2019, 28), Firth advocated a type of "polysystemic" linguistic analysis that was interested in different, congruent modes of meaning, whether phonetic, phonological, syntactic, or semantic, but always situated in broader social context (Firth 1957c, 27, 30 ; Robins 1997, 214). In stark contrast to Harris, Firth explicitly rejected any efforts to create "unity in linguistics" (Firth, 1968d, 48) or one system of analysis.

[4] An important early figure in NLP, Sager received her PhD in Linguistics at the University of Pennsylvania and was directly influenced by Harris's work. See, e.g, Hirschman, Grishman, and Sager (1975).

Citing the later Wittgenstein (1953, Firth 1957c, 11), Firth was mainly interested in the concrete use of language, reversing the schema of Ferdinand de Saussure (1916/2011) in which language (*la langue*) is a system "external to and on a different plane from individual phenonema," including the concrete instances (*la parole*) of language use (Firth 1949, 400; Firth 1950, 44–45). While his final ideas matured considerably later, Firth initially articulated his ideas about semantics in two 1935 papers, one on semantics (Firth, 1935a) and one on phonology (Firth, 1935b), using the term "contextual distribution" in both. However, Firth (1957c, 18) ultimately disavowed this initial distributional theory—which was not too dissimilar from Harris's—as "useful" but inadequate to act as the "main principle" in a theory "of structures involving the statement of the values of the elements of structure by reference to systems."

By distinguishing between system (syntagm) and structure (paradigm), Firth wanted to highlight two operational principles necessary for meaning by collocation: 1) substitution within "the same level of abstraction," and 2) commutation across different levels (Robins, 1953, 140). Only substitution that does not produce commutation in a sequence, indicates similarity of value or function (Firth 1957c, 5; Firth 1968c, 23)[5]. Two words are only substitutable—and hence similar in function and meaning—if their values do not commute across a particular sentence. Substitutability, then, does not equal synonymy. Take, for example, the following two phrases containing a) prepositional and b) adverbial uses of the word "by":

(a) They go by night.
(b) They go by night after night.

Now, "by" could be replaced by the word "past" without commuting the meaning of the other words in (b). However, replacing "by" with "past" would commute with the rest of (a) in an impossible way (Firth, 1968c, 23–24). This demonstrates how substitution concerns the relationship between "by" and "past" as two elements at the same level of analysis—i.e., lexical units—but in order to account for commutation, we need to look beyond this level to other levels of abstraction.[6]

Firth's conception of collocation and his frequent nods to Whitehead were part of his "monistic" approach that rejected the division between mind and body (Firth 1957c, 2; Palmer 1968, 5) and all the other dualities—language and thought, word and idea, signifier and signified, expression and content (Firth, 1951, 86)—that characterized the structuralist linguistics of his time. He similarly rejected any notion of linguistics as "a *theory of universals* for *general linguistic description*" (Firth, 1957c, 21). Anticipating contemporary concerns about language diversity in NLP (e.g. Bender, 2019), Firth called for the Western scholar to "de-Europeanize himself" and the English scholar, due to the universal use of his language, to "de-Anglicize himself" (Firth 1968a, 96; Senis 2015, 274).

4.2 Context and connection

Diverging from structuralist linguistics, Firth suggested that a text should always be given a "renewal of connection with experience" (Firth, 1957c, 29). This notion of meaning was influenced by Malinowski, for whom Firth worked as an assistant early in his career (Plug, 2008, 346) and from whom he borrowed the notion of "context of situation" (Firth 1935a, Robins 1997, 211). In Malinwoski's view, meaning was more than just a dyadic relationship between a word and its referent, "a multidimensional and functional set of relations between the word in its sentence and the context of its occurrence" (Robins, 1971, 35). However, while Malinowski's view on meaning was entirely functional and hyper-local, Firth employed the notion of "context of situation" as a necessary abstraction, not as a shorthand for things in themselves (Firth 1950, 43, Palmer 1968, 6). Context of situation is derived from an analytical choice, "a set of categories in ordered relations abstracted from the life of man in the flux of events, from personality in society," (Firth, 1957c, 30) prehending something of importance and bracketing the rest. It is, then, not necessarily about restricting the meaning of every utterance to a specific time and place, but about defining "an abstract set of semantically relevant categories, abstracted from multitudes of actual situations, to which unique particulars could be referred." (Robins, 1971, 41–42) Firth called for the linguist to focus on "attested language text duly recorded", accounting for a text's associated context of situation *and* its interior relations. (Firth,

[5] For further details, see examples provided by Bursill-Hall (1960).

[6] A useful analogy might be the way in which BERT handles different aspects of language at different layers of the model (Tenney et al., 2019, e.g.). However, no matter how large, a language model like BERT does not account for the context of situation.

1957c, 29–30)

Firth was famously opaque with the exact operationalization of his concepts, including context of situation, but he did provide a detailed list of the different contextual elements that a linguist should bring into relation during analysis (Firth, 1950, 43). These include the relevant features of participants (persons, personalities); their verbal and non-verbal actions; the relevant objects; and the effects of verbal action.

During Firth's lifetime, the most thorough work that put his notion of context of situation to work was an ethnographic study by his student T.F. Mitchell in former Cyreneica (today Libya) on the language of buying and selling at the local markets of different cities and villages in the region. For Mitchell (1957, 32–33), contexts that might "correlate" with particular types of text included: the spatio-temporal situation of persons in the context; the activities of participants; the attitudes of the participants; their "personalities" such as specific trade of profession, geographical and class origins, educational standard, inter-relationship, and so on.

It is worth noting that both Mitchell and M.A.K. Halliday—Firth's student who synthesized much of his theory—used words such as "correlation," "inference," and "prediction" to describe the relationship between a text and its situational context, implying that a statistical extension of their approaches would not be completely unreasonable. In fact, Halliday himself suggested as much, when he in the early 1990s made efforts to bridge his branch of linguistics with the nascent field of corpus linguistics (Halliday, 1991).

In conclusion: Firth's famous quote itself refers to collocation, while his notion of "context" implies something much broader, "the whole conceptual meaning" (Firth, 1957c, 11). Context is the ground against which the figure of the text must be understood, no matter (per Harris) how "detached" its pattern might be (e.g. Auer, 1996). Without context, collocation captures only one narrow "mode of meaning."

4.3 Restricted languages

Like Harris, Firth's revival in connection with 1990s corpus linguistics was related to his attempts to respond to practical needs of empirical research. Expanding upon his functional understanding of language, Firth developed his notion of "restricted languages" in the 1950s (Léon 2007, 7). In a posthumously published essay, he describes social actors as "collect[ing] a varied repertory of interlocking roles" corresponding to a "constellation of restricted languages" (Firth, 1968e, 207). As people shift between locally contextualized roles, they draw upon their "repertory" of restricted languages with specialized vocabulary and discursive styles that both reflect and constitute these contexts. Thus one might speak of a "restricted language of science, sport, defense, industry, aviation, military services, commerce, law and civil administration, politics, literature, etc." (Léon, 2008, 261). As such, the concept of restricted language is now generally seen as a precursor to the concept of "register," which was taken up by subsequent sociolinguists and linguistic anthropologists (e.g. Halliday, 1968; Gumperz and Hymes, 1972; Agha, 2005).

In proposing restricted languages as the proper object of descriptive linguistic analysis, Firth was making a broader theoretical point against, on the one hand, "the monosystemic view of language" of neo-Bloomfieldians like Harris and "pointless discussions on metalanguage" on the other, for metalanguage could be reanalyzed as a "restricted language of linguistics" itself (Léon, 2007, 9). Simply put, a descriptive linguistics which privileges restricted languages also necessarily privileges contexts of situation as an essential dimension of variation that allows social meaning to inhere in language.

5 Discussion: Words in mixed company

Often cited, together or separately, to justify a distributional appraoch to semantics, Firth and Harris nonetheless offer differing views on language and meaning. Harris offers us a rigorous formalism that treats language as a "detached pattern"—not a "code", but a particular system of meaning. Firth, by comparison, left a much more scattered legacy that was only systematized by his students. Firth and Harris shared a concern about the lack of an external metalanguage of linguistics, but drew different conclusions from it. If Harris responded to this conundrum by creating one hierarchically organized system without intrusion from extra-linguistic factors, Firth called for an investigation of language as a "spectrum" (Firth, 1951, 76) with different modes of meaning that had to be addressed through multiple levels of analysis—starting with the context of situation and proceeding from there to decide which other levels are

relevant (Firth, 1950, 44). Firth's distributional theory has been described unfavourably as based on frequent co-occurrence, in contrast to the recursive dependencies developed by Harris (Habert and Zweigenbaum, 2002, 205). For Harris, the meaning of a word depends on its set (e.g. Harris, 1991, 17) such that, for example, the words "divide" and "multiply" operate on the word "cell" (and vice versa) in the same way, producing essentially the same meaning (Harris 1988, 62). However, Firth's final method of substitution and commutation also establishes complex, multidimensional criteria for distributional contrast as well as a framework for understanding polysemy. Though less formalized and less obviously recursive than Harris', Firth's approach can, arguably, also be read as treating linguistic elements as operators and arguments defined by their sets in a complex hierarchy (Firth, 1950, 44; Firth, 1951, 76 "at a series of congruent levels" (Firth, 1957c, 29–30) with different "bands of abstraction" (Firth, 1968d, 49), *including* the extra-linguistic context of situation. Harris and Firth both understand any linguistic analysis as incomplete, for Harris always a pursuit of the least description" (Harris, 1988, 3)—i.e., best "grammar" or model—for Firth always grounded in the social construction of facts, without any possibility of "complete axiomatization" (Firth, 1968d, 44–45).

From the perspective of empirical work, especially decades after their time, Firth and Harris also share similarities. Both rejected the mentalism that was so prevalent during their time. They were both revived as empiricist originators during the rise of statistical learning in the 1990s, and their respective work on restricted languages and sublanguages largely conflated in service of the practical concerns of corpus linguistics (Léon, 2008). Their theories both included in what we might call a "reality principle," a final arbiter of meaning outside of form: the pair test for Harris, the context of situation for Firth. The former grounds linguistics in the smallest possible unit of analysis as understood by the native speaker, the latter in social actions and objects.

In light of recent calls to extend the "world scope" (Bisk et al., 2020) of NLP and to move towards pragmatic notions of meaning, it might make sense to balance Harris's formalism and Firth's pluralism. Though Firth warned us against overextending linguistics, he was generous with the company that words could keep. They mingled with each other, but also with events, objects, people, and indexical features such as time and space. And if NLP is ready to move beyond the corpus, then even Harris might acknowledge that when modeling language in "the immediate situation"—whether in online interactions or face-to-face communication—words are *not* on their own, that to judge the meaning of a combination of words, we can summon "the aid of some of the extra-linguistic situational information" (Harris, 1991, 368). In the following subsections, we consider two ways in which NLP is already doing this, in order to highlight some already existing strategies for broader contextualization. We call these strategies "comparative stratification" and "syntagmatic extension."

5.1 Comparative stratification

Corpus linguistics emerges from the question of what kind of company words keep, depending on their context. The issue of context was motivated by the introduction of corpus linguistics both for students of Firth—who considered restricted languages as a way of handling context—and for Harris's sublanguages, which were "contextually situated and suitable for being processed automatically" (Léon, 2021, 149–150). However, beyond just studying the restricted corpus, we might also consider the ways in which large datasets can be "stratified," systematically dividing them into subcorpora that are studied in relation to each other. Here, the company that words keep among each other is *limited* for analytical purposes, but in a manner that implies a relationship between that "company" and situational context.

Diachronic embeddings are especially representative of this approach. By stratifying timestamped data into a number of intervals, training separate models for each and then aligning the embeddings using either "second-order embeddings" or methods such as linear transformations (Kutuzov et al., 2018), the analyst can effectively represent a temporal "context of situation." This approach works with both static (Hamilton et al., 2016) and contextual embeddings (Martinc et al., 2020). While variants of this approach are most commonly used to study semantic shifts (e.g., Garg et al., 2018; Kozlowski et al., 2019; Mendelsohn et al., 2020; Brunila and LaViolette, 2021), it could plausibly be used to stratify a dataset according to other variables such as space (e.g. Bamman et al., 2014; Gong et al., 2020), online communities (Lucy and

Bamman, 2021), persons (Yao et al., 2020), or domains (Spinde et al., 2021). Words still only keep the company of one another, but by limiting their company we implicitly introduce other participants in the analysis.[7]

5.2 Syntagmatic extension

Recall that for Firth, meaning by collocation and considerations of a "typical context of situation" (Firth, 1950, 44) were exercises in abstraction, with collocation being an abstraction at the syntagmatic level. Instead of restricting the company words keep, we might follow Firth's recommendation to consider them in wider company "of the same abstract nature" (Firth, 1950, 7). In vector semantics, this would imply that we explicitly introduce different contextual factors in the *same* vector space with our words, endowing them all with ontological equality.

The paragraph vectors introduced by Le and Mikolov (2014) as an extension of the earlier Skip-gram algorithm (Mikolov et al., 2013) are representative for this approach. In practice, this method extends the syntagmatic chain of words by introducing a vector for the document as a new paradigmatic element. In principle, this type of "global context" (Grbovic and Cheng, 2018) could be anything and include several paradigmatic elements, as we can see in the research on multi-modal embeddings (e.g. Baroni, 2016) and generative modeling (e.g. Ramesh et al., 2022).[8] Models have been developed that include demographic (Garimella et al., 2017) or persona (Li et al., 2016) vectors in the embedding space, such that intra-textual relations are accompanied by information about speakers' social categories. However, implemented with static embeddings and without some additional grammar restrictions, these context vectors essentially add only a "bag of contexts." For static embeddings, additional grammar constraints could be introduced, as was done in research on Point-of-Interest (POI) data in the the domain of geosemantics, where researchers constrained contextual vectors using spatial variograms (Yan et al., 2017). Beyond static embeddings, large-language models and their dynamic embeddings could either be pretrained (with the appropriate dataset) or finetuned on data with text associated with different contextual variables. This would realize the proposal that Halliday made in the early 1990s when he suggested an extension of the language modeling schema from the early work of Shannon, to a model with "global probabilities, those of the grammar of English, and the locally conditioned probabilities, those of this or that particular register" (Halliday, 1991, 37).

6 Conclusions

This paper revisited the theories of the two most well-known progenitors of the distributional approach to meaning in NLP. Recognizing the open question of how to bring NLP beyond the corpus, we offer a thorough account of the two distributional theories that are most often invoked to justify the modeling of meaning through departures from randomness in the company that words keep. Comparing the work of Harris and Firth—who both published their major work before the rise of the internet and its corpora—we find two distinct theories of distribution: one formal and mathematical, treating language as a particular type of detached information, another more schematic and anthropological, treating language as a functional spectrum which always emanates from a particular context of situation. The legacies of both Firth and Harris can be seen in the current paradigm of corpus linguistics, but in the domain of distributional semantics, it is Harris's ethos that dominates, despite Firth providing its most famous tagline.

Moving forward, we suggest that semantic modeling take more inspiration from Firth, and consider the context of situation and the wide variety of company that words can keep as crucial sites of innovation for the field. Doing so may not involve following a finite set of steps or flowchart. Rather, we humbly suggest that the field may be enriched by thoughtful and creative re-engagements with the intellectual traditions from which it has historically drawn. This does not imply abandoning the rigor provided by Harris. On the contrary, we find that Firth and Harris would probably have agreed that any model or "grammar" is inevitably incomplete and partial. No universal model is possible, despite the large-language modeling fervor, nor will there be one theory of language to guide us. There are only the partial perspectives and the inevitable choice of adapting one.

[7] In a very broad sense, the trend of pretraining large language models and then finetuning them on specific datasets is of course also an admission of the importance of "context of situation."

[8] Firth himself (1957c, 26) recommended accompanying word definitions and collocational information with pictures.

Acknolwedgements

We would like to thank Dan Jurafsky and Jacqueline Léon for their gracious and helpful comments which greatly improved the paper. Mikael Brunila is funded by the Kone Foundation, and Jack LaViolette is funded by the Mellon Foundation. We thank them for their generous support.

References

Asif Agha. 2005. Voice, footing, enregisterment. *Journal of Linguistic Anthropology*, 15(1):38–59.

Peter Auer. 1996. From context to contextualization. *Links & Letters*, 3(1):11–28.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, May 7-9, 2015, Conference Track Proceedings*, pages 1–15, San Diego, CA, USA.

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, MD, USA. Association for Computational Linguistics.

Marco Baroni. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Emily M. Bender. 2019. The #BenderRule: On naming the languages we study and why it matters. *The Gradient*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2002. Latent Dirichlet Allocation. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 601–608. MIT Press.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Pierre Bourdieu. 1984. *Distinction: A social critique of the judgement of taste*. Harvard University Press, Cambridge, MA.

E. Bruni, N. K. Tran, and M. Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Mikael Brunila and Jack LaViolette. 2021. WMDecompose: A Framework for Leveraging the Interpretable Properties of Word Mover's Distance in Sociocultural Analysis. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 154–167, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Levi R. Bryant. 2011. *The democracy of objects*. Open Humanities Press, Ann Arbor, MI, USA.

Curt Burgess. 1998. From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, 30(2):188–198.

G. L. Bursill-Hall. 1960. Levels analysis: J. R. Firth's theories of linguistic analysis. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 6(2):124–135.

David G. Butt. 2001. Firth, Halliday and the development of systemic functional theory. In Sylvain Auroux, E. F. K. Koerner, Hans-Josef Niederehe, and Kees Versteegh, editors, *History of the language sciences*, volume 2 of *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)*, pages 1806–1838. De Gruyter, Berlin.

David G. Butt. 2013. Whiteheadian and functional linguistics. In Michael Weber, editor, *Handbook of Whiteheadian Process Thought*, pages 21–32. De Gruyter, Berlin.

David G. Butt. 2019. Firth and the origins of systemic functional linguistics: Process, pragma, and polysystem. In David Schönthal, Geoff Thompson, Lise Fontaine, and Wendy L. Bowcher, editors, *The Cambridge Handbook of Systemic Functional Linguistics*, Cambridge Handbooks in Language and Linguistics, pages 11–34. Cambridge University Press, Cambridge, UK.

Noam Chomsky. 1972. *Language and mind*. Harcourt Brace Jovanovich, New York, NY, USA.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *arXiv:2204.02311 [cs]*. ArXiv: 2204.02311.

William A. Christian. 1959. *An interpretation of Whitehead's metaphysics*. Yale University Press, New Haven, CT, USA.

Kenneth Church and Mark Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4:1–18.

Kenneth W. Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Kenneth W. Church and Robert L. Mercer. 1993. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics*, 19(1):1–24.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Ferdinand de Saussure. 1916/2011. *Course in general linguistics*. Columbia University Press, New York, NY, USA.

Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Richard A. Harshman, Thomas K. Landauer, Karen E. Lochbaum, and Lynn A. Streeter. 1989. Computer information retrieval using latent semantic structure. US patent US4839853A.

Scott C. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Gilles Deleuze and Félix Guattari. 1987. *A thousand plateaus: Capitalism and Schizophrenia*. University of Minnesota Press, Minneapolis, MN, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41(1):87–100.

Jacob Eisenstein. 2019. *Introduction to natural language processing*. MIT Press, Cambridge, MA, USA.

Robert M. Fano. 1961. *Transmission of information: A statistical theory of communication*. MIT Press, Cambridge, MA, USA.

J.R. Firth. 1935a. The technique of semantics. *Transactions of the Philological Society*, 34(1):36–73.

J.R. Firth. 1935b. The use and distribution of certain English sounds. *English Studies*, 17(1-6):8–18.

J.R. Firth. 1949. The semantics of linguistic science. *Lingua*, 1:393–404.

J.R. Firth. 1950. Personality and language in society. *The Sociological Review*, a42(1):37–52.

J.R. Firth. 1951. General linguistics and descriptive grammar. *Transactions of the Philological Society*, 50(1):69–87.

J.R. Firth. 1957a. Introduction. In J.R. Firth, editor, *Studies in Linguistic Analysis*. Basil Blackwell, Oxford, UK.

J.R. Firth. 1957b. Modes of meaning. In *Papers in Linguistics, 1934-1951*, pages 190–215. Oxford University Press, London, UK.

J.R. Firth. 1957c. A synopsis of linguistic theory, 1930-1955. In J.R. Firth, editor, *Studies in Linguistic Analysis*. Basil Blackwell, Oxford, UK.

J.R. Firth. 1968a. Descriptive linguistics and the study of English. In Frank R. Palmer, editor, *Selected papers of J.R. Firth (1952-59)*, pages 96–113. Longman and Indiana University Press, London, UK, and Bloomington, IN, USA.

J.R. Firth. 1968b. Ethnographic analysis and language with reference to Malinowski's views. In Frank R. Palmer, editor, *Selected papers of J.R. Firth (1952-59)*, pages 137–168. Longman and Indiana University Press, London, UK, and Bloomington, IN, USA.

J.R. Firth. 1968c. Linguistic analysis as a study of meaning. In Frank R. Palmer, editor, *Selected papers of J.R. Firth (1952-59)*, pages 12–27. Longman and Indiana University Press, London, UK, and Bloomington, IN, USA.

J.R. Firth. 1968d. Structural linguistics. In Frank R. Palmer, editor, *Selected papers of J.R. Firth (1952-59)*, pages 35–53. Longman and Indiana University Press, London, UK, and Bloomington, IN, USA.

J.R. Firth. 1968e. The treatment of language in general linguistics. In Frank R. Palmer, editor, *Selected papers of J.R. Firth (1952-59)*, pages 206–209. Longman and Indiana University Press, London, UK, and Bloomington, IN, USA.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):E3635–E3644. ISBN: 9781720347118 publisher: National Academy of Sciences section: PNAS Plus.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295.

Arthur M. Glenberg and David A. Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 43(3):379–401.

Yoav Goldberg. 2017. Neural network models for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

John A. Goldsmith. 2005. The legacy of Zellig Harris: Language and information into the 21st century, vol. 1: Philosophy of science, syntax and semantics (review). *Language*, 81(3):719–736.

Hongyu Gong, Suma Bhat, and Pramod Viswanath. 2020. Enriching Word Embeddings with Temporal and Spatial Information. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 1–11, Online. Association for Computational Linguistics.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. MIT Press, Cambridge, MA, USA.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.

Mihajlo Grbovic and Haibin Cheng. 2018. Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, pages 311–320, London, United Kingdom. ACM Press.

John Joseph Gumperz and Dell Hymes, editors. 1972. *Directions in sociolinguistics: The ethnography of communication*. Holt, Rinehart and Winston, New York, NY, USA.

Benoît Habert and Pierre Zweigenbaum. 2002. Contextual acquisition of information categories - What has been done and what can be done automatically? In Bruce E. Nevin and Stephen B. Johnson, editors, *The Legacy of Zellig Harris - Language and information into the 21st century*, volume 2: Mathematics and computability of language. John Benjamins Publishing Company.

A. Halevy, P. Norvig, and F. Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12.

M.A.K. Halliday. 1968. The users and uses of language. In Joshua A. Fishman, editor, *Readings in the sociology of language*, pages 139–169. De Gruyter Mouton, New York, NY.

M.A.K. Halliday. 1991. Corpus studies and probabilistic grammar. In Karin Aijmer and Bengt Altenberg, editors, *English Corpus Linguistics*, pages 30–43. Routledge, London, UK.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Zellig S. Harris. 1941. Review of Grundzüge der Phonologie. *Language*, 17(4):345–349.

Zellig S. Harris. 1951. *Methods in structural linguistics*. University of Chicago Press, Chicago, IL, USA.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Zellig S. Harris. 1968. *Mathematical structures of language*. John Wiley & Sons, New York, NY, USA.

Zellig S. Harris. 1988. *Language and information*. Columbia University Press, New York, NY, USA.

Zellig S. Harris. 1991. *A theory of language and information: A mathematical approach*. Clarendon Press, Oxford, UK.

Zellig S. Harris, Michael Gottfried, Thomas Ryckman, Anne Daladier, and Paul Mattick. 1988. *The form of information in science: Analysis of an immunology sublanguage*. Kluwer Academic Publishers, Amsterdam.

James Henderson. 2020. The unstoppable rise of computational linguistics in deep learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6294–6306, Online. Association for Computational Linguistics.

Donald Hindle. 1990. Noun classification from predicate-argument structures. In *28th Annual meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, PA, USA. Association for Computational Linguistics.

Lynette Hirschman, Ralph Grishman, and Naomi Sager. 1975. Grammatically-based automatic word class formation. *Information Processing & Management*, 11(1-2):39–57.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy. 2018. The social and the neural network: How to make natural language processing about people again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 42–49, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Stephen B. Johnson. 2002. The Computability of Operator Grammar. In Bruce E. Nevin and Stephen B. Johnson, editors, *The Legacy of Zellig Harris - Language and information into the 21st century*, volume 2: Mathematics and computability of language. John Benjamins Publishing Company. Publication Title: cilt.228.

Martin Joos. 1950. Description of language design. *The Journal of the Acoustical Society of America*, 22(6):701–707.

Dan Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, Hoboken, NJ, USA.

Dan Jurafsky and James H. Martin. 2021. *Speech and language processing*, 3rd (draft) edition.

Richard Kittredge and John Lehrberger. 1982. *Sublanguage: Studies of language in restricted semantic domains*. de Gruyter, Berlin.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

William Labov. 1966/2006. *The social stratification of English in New York City*, 2nd edition. Cambridge University Press, Cambridge, UK.

William Labov. 1972. *Sociolinguistic patterns*. University of Pennsylvania press, Philadelphia, PA, USA.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196. PMLR.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, MD, USA. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from wor embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 994–1003.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In Vivek Nigam, Tajana Ban Kirigin, Carolyn Talcott, Joshua Guttman, Stepan Kuznetsov, Boon Thau Loo, and Mitsuhiro Okada, editors, *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, Lecture Notes in Computer Science, pages 189–202. Springer International Publishing, New York, NY, USA.

Li Lucy and David Bamman. 2021. Characterizing english variation across social media communities with bert. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Jacqueline Léon. 2007. From linguistic events and restricted languages to registers: Firthian legacy and corpus linguistics. *The Henry Sweet Society Bulletin*, 49(1):5–25.

Jacqueline Léon. 2008. Empirical Traditions of Computer-Based Methods. Firth's Restricted Languages and Harris' Sublanguages. *Beiträge zur Geschichte der Sprachwissenschaft*, 18(2):259.

Jacqueline Léon. 2011. Z. S. Harris and the semantic turn of mathematical information theory. *History of Linguistics 2008*, pages 449–458. Publisher: John Benjamins.

Jacqueline Léon. 2021. *Automating linguistics*. Springer Nature, London, UK.

Christopher D. Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Grant McKenzie and Benjamin Adams. 2021. Natural language processing in GIScience applications. In John P. Wilson, editor, *Geographic Information Science & Technology Body of Knowledge*, 4th quarter 2021 edition edition.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3:1–24.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., Redhook, NY, USA.

T. F Mitchell. 1957. The language of buying and selling in Cyrenaica: A situational statement. *Hespéris - Archives Berbères et Bulletin d'Institut des Hautes Études Marocaines*, 44:31–71.

Bruce Nevin. 2010. Noam and Zellig. In Douglas A. Kibbee, editor, *Chomskyan (R)evolutions*, pages 103–168. John Benjamins Philadelphia.

Bruce E. Nevin. 1993. A minimalist program for linguistics: The work of Zellig Harris on meaning and information. *Historiographia Linguistica*, 20(2-3):355–398.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.

Peter Norvig. 2012. Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance*, 9(4):30–33.

Olasope O. Oyelaran. 1967. Aspects of linguistic theory in Firthian linguistics. *WORD*, 23(1-3):428–452.

F.R. Palmer. 1968. Introduction. In J.R. Firth, editor, *Selected Papers of J.R. Firth, 1952-59*. Longmans, Green and Co., London, UK.

P. B. Pandit. 1970. Review of Selected Papers of J. R. Firth 1952-59. *Journal of Linguistics*, 6(2):280–284. Publisher: Cambridge University Press.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, LA, USA. Association for Computational Linguistics.

Leendert Plug. 2008. J. R. Firth: a new biography. *Transactions of the Philological Society*, 106(3):337–374.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 759–766, New York, NY, USA. Association for Computing Machinery.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *OpenAI*, pages 1–26.

R. H. Robins. 1953. The phonology of the nasalized verbal forms in sundanese. *Bulletin of the School of Oriental and African Studies, University of London*, 15(1):138–145.

R. H. Robins. 1971. Malinowski, Firth, and the "context of situation". In Edwin Ardener, editor, *Social Anthropology and Language*. Routledge, London, UK.

R. H. Robins. 1997. The contribution of John Rupert Firth to linguistics in the first fifty years of *Lingua*. *Lingua*, 100(1):205–222.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Stuart J. Russell and Peter Norvig. 2020. *Artificial intelligence: A modern approach*, 4th edition. Pearson, Hoboken, NJ, USA.

Naomi Sager. 1975. Sublanguage grammers in science information processing. *Journal of the American Society for Information Science*, 26(1):10–16.

Naomi Sager. 1981. *Natural language information processing: A computer grammar of English and its applications*. Addison-Wesley Publishing Company, Advanced Book Program, Boston, MA, USA.

Magnus Sahlgren. 2008. The distributional hypothesis. *Rivista di Linguistica*, (20.1):33–53.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implcations of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Hinrich Schütze and Jan Pedersen. 1993. A vector model for syntagmatic and paradigmatic relatedness. In *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research*, pages 104–113, Oxford, UK.

Hinrich Schütze. 1992. Dimensions of meaning. In *Supercomputing '92: Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 787–796. IEEE Computer Society, Washington, DC, USA.

Hinrich Schütze. 1993. Word space. In *Advances in Neural Information Processing Systems*, volume 5, pages 895–902. Morgan-Kaufmann, San Francisco, CA, USA.

Angela Senis. 2015. The contribution of John Rupert Firth to the history of linguistics and the rejection of the phoneme theory. In *ConSOLE XXIII (23rd Conference of the Student Organization of Linguistics in Europe)*, pages 273–293, Paris, France. Leiden University Centre for Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656.

C.E. Shannon. 1945. A mathematical theory of cryptography - Case 10878. Technical report, Bell Telephone Laboratories, Princeton Libraries.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, 23(3-4):193–229.

Timo Spinde, Lada Rudnitckaia, Felix Hamborg, and Bela Gipp. 2021. Identification of biased terms in news articles by comparison of outlet-specific word embeddings. In *International Conference on Information*, pages 215–224. Springer.

Ronen Tamari, Chen Shani, Tom Hope, Miriam R L Petruck, Omri Abend, and Dafna Shahaf. 2020. Language (re)modelling: Towards embodied language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6268–6281, Online. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of the 2019 International Conference on Learning Representations*, pages 1–17, New Orleans, LA, USA.

Margaret Thomas. 2011. *Fifty key thinkers on language and linguistics*. Routledge, London, UK.

Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. 2020. (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170–5184, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:6000–6010.

Warren Weaver. 1952. Translation. In *Proceedings of the Conference on Mechanical Translation*, pages 1–12, Cambridge, MA, USA.

Alfred North Whitehead. 1938. *Modes of thought*. Cambridge University Press, Cambridge, UK.

Alfred North Whitehead. 1957. *Process and Reality*. Macmillan, New York, NY.

Ludwig Wittgenstein. 1953. *Philosophical investigations*. Macmillan, New York, NY, USA.

Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. 2017. From ITDL to Place2Vec: Reasoning about place type similarity and relatedness by learning embeddings from augmenteed spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'17*, pages 1–10, Redondo Beach, CA, USA. ACM Press.

Jing Yao, Zhicheng Dou, and Ji-Rong Wen. 2020. Employing personal word embeddings for personalized search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1359–1368.